AD-777 175

INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE. SECOND LANGUAGE LEARNING

Daniel N. Kalikow, et al

Bolt, Beranek and Newman, Incorporated

Prepared for:

Air Force Office of Scientific Research
Advanced Research Projects Agency

31 December 1973

INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE

TECHNICAL REPORT

SECOND LANGUAGE LEARNING

31 December 1973

by

Daniel N. Kalikow
and
Ann M. Rollins

ARPA Order No. 1993

Sponsored by the Advanced Research Projects Agency,
Department of Defense, under Air Force Office of
Scientific Research Contract F44620-71-C-0065

Prepared for

Air Force Office of Scientific Research
1400 Wilson Boulevard
Arlington, Virginia 22209

ib

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR - TR - 73 - 2334	2. GOVT ACCESSION NO.
	3. RECIPIENT'S CATALOG NUMBER AD 777175
4. TITLE (and Subtitle) INFORMATION PROCESSING MODELS AND COMPUTER AIDS FOR HUMAN PERFORMANCE	5. TYPE OF REPORT & PERIOD COVERED Interim
	6. PERFORMING ORG. REPORT NUMBER 2654
7. AUTHOR(s) Daniel N. Kalikow Ann M. Rollins	8. CONTRACT OR GRANT NUMBER(s) F44620-71-C-0065
9. PERFORMING ORGANIZATION NAME AND ADDRESS Bolt, Beranek and Newman, Inc. 50 Moulton Street Cambridge, Massachusetts 02138	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61101E 1993-04 681313
11. CONTROLLING OFFICE NAME AND ADDRESS Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia 22209	12. REPORT DATE 31 December 1973
	13. NUMBER OF PAGES 58
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Air Force Office of Scientific Research (NL) 1400 Wilson Boulevard Arlington, Virginia 22209	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release;
distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The task is to carry out the final development of a computer-based system for automated instruction of the new speech sounds of second languages, and to field-test this system for two language pairs: English speakers learning Mandarin Chinese, and Spanish speakers learning English. This report describes the first evaluation experiment of the Mark II model of the Automated Pronunciation Instructor (API) system. Two matched groups of students of Elementary Mandarin Chinese were studied. One group was tested and trained with the API system;

20. ABSTRACT (Continued)

the other was simply tested within the same time frame. Significant treatment effects were observed.

TABLE OF CONTENTS

TECHNICAL REPORT

31 December 1973

ARPA Order No. 1993

Program Code No. 3D20

Contractor: Bolt Beranek and Newman Inc.

Effective Date of Contract: 1 January 1971

Contract Expiration Date: 30 June 1974

Amount of Contract: $689,260

Contract No. F44620-71-C-0065

Principal Investigators: John A. Swets
Daniel N. Kalikow

Telephone No. 617-491-1850

Title: Information Processing Models and
Computer Aids for Human Performance

SUMMARY

1. Technical Problem

The task is to carry out the final development of a computer-based system for automated instruction of the new speech sounds of second languages, and to field-test this system for two language pairs: English speakers learning Mandarin Chinese, and Spanish speakers learning English.

2. General Methodology

Laboratory experiments and field evaluations.

3. Technical Results

This report describes the first evaluation experiment of the Mark II model of the Automated Pronunciation Instructor (API) system. Two matched groups of students of Elementary Mandarin Chinese, enrolled at two local universities, were studied. One group was tested and trained with the API system; the other was simply tested within the same time frame. Significant treatment effects were observed.

4. Department of Defense Implications

Language schools of the Department of Defense give instruction in approximately 65 languages to over 200,000 students each year. The systems under development are designed to facilitate this instructional process.

LIST OF TABLES

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contributions of the other members of the project team: Douglas W. Dodds, Kenneth N. Stevens, and John A. Swets. The advice of Raymond S. Nickerson is greatly appreciated. Thanks are also due to Victor Zue of the staff of M.I.T.'s Lincoln Laboratory for his assistance and comments on the Chinese curriculum, and to the staff of the Department of Far Eastern Languages of Harvard University, in particular Drs. Edward Wagner, Rulan Pian, and Shou-ying Lin.

The advice of Drs. Austir Kibler and George Lawrence of ARPA, and of Drs. Glen Finch and Charles Hutchinson of AFOSR is also acknowledged.

PREFACE

The presen: contract is a partial continuation of a research program begun in 1966 under ARPA sponsorship. Of the four tasks at one time funded under AFOSR Contract F44620-67-C-0033, the present task remains active under Contract F44620-71-C-0065. This technical report covers the period extending through 31 December 1973, and is devoted to a description of experimental activities completed earlier in that calendar year. It completes the description of the first phase of the final testing of the Automated Pronunciation Instructor (API) system, in one of two language pairs: English speakers learning Mandarin Chinese pronunciation. The second evaluation, currently proceeding on schedule at the University of Miami, Coral Gables, Florida, is much more extensive. It involves Spanish speakers learning English pronunciation. That field test will be the subject of future reports.

1. INTRODUCTION

The purpose of the present experiment is the evaluation of the effectiveness of the Automated Pronunciation Instructor (API) system in the modification of the speech of English-speaking students of Mandarin Chinese. The design concepts of the API have been detailed in previous technical reports, but a brief sketch of the system and its operation in the context of the English-Chinese language pair is presented here as a prelude to the description of the experiment undertaken.

1.1 Background

The central problem to which the API system addresses itself is that students of new languages bring to their effort certain pronunciation handicaps forced on them by their overlearned skill in their "mother tongue." The distinguishing factor of the API approach is its production of visual as well as auditory correlates of the utterances of both student and teacher. By intelligent and interactive use of this double-modality feedback, the student's pronunciation may be improved in a manner unavailable to the student using audio feedback alone. The relative inefficiency of the audio channel arises because of the nature of the second-language learning task: certain sound distinctions that are phonemic in the target language are, by coincidence, not present or open to free variation in the source language; and the resultant inability of the student to perceive or to produce those distinctions both defines and circumscribes the parameters of his accent.

The API system deals with this problem by concentrating the efforts of the student within those sound distinctions known by contrastive language analysis to be major contributors to the overall accent he exhibits. The predictability and generality of the problems across many students of similar background and target-language objective (referred to as students of a given "language pair") makes possible a group approach. At present, technical constraints have resulted in a system that handles but one student at a time, but expansion to a multi-station configuration is a feasible later goal, if warranted. The evaluation experiments reported here have been carried out with groups of students using the API system on a staggered-schedule basis.

1.2 A Brief Sketch of the API System

The API system is built around a minicomputer (Digital Equipment Corporation PDP-8e) which the student controls by means of a few pushbuttons. It is actually easier for the student to manipulate this system than the equipment in a conventional language laboratory with facilities for recording and playback of student and teacher speech. The API contains those features and adds to them a real-time visual analysis of certain aspects of his speech.

The visual display is produced in such a way as to accentuate the expected differences between his and the teacher's rendition of a selected set of training utterances. Through an understanding of the relationships between visual display and the manner of articulation, the student is guided towards articulatory gestures more closely approximating the teacher's.

The student wears a headband-mounted microphone, positioned close to the mouth but out of the breath stream. He also wears a miniature accelerometer, fastened to the throat with thin double-surfaced adhesive tape. This transducer picks up the fundamental frequency, or "tone," of the voice (i.e., the rate at which the vocal cords are vibrating during voiced portions of speech). The microphone-accelerometer assembly is comfortable for the student, who quickly forgets its presence and concentrates on the task at hand.

The student receives feedback from a large display oscilloscope and a high fidelity loudspeaker. The computer draws pictures on the screen while performing its other chores of controlling data input, storage and the rest of the equipment of the system. Descriptions of the displays themselves will be given below in the context of the curriculum.

The student informs the system which of several operations he wishes to perform through the use of pushbuttons recessed within his work table. There are buttons for recording, playback, display manipulation, new training utterances, and other utility functions.

At no time during the operation of the system does the equipment ever make an evaluation of the adequacy of the pronunciation of the student. That is left to the student, on the hypothesis that the additional information provided by the visual analysis, in conjunction with the audio replay, will suffice to bring the student's abilities as a pattern recognizer into play.

1.3 Phonological Contrasts in the English-Mandarin Chinese Language Pair

Two major pronunciation problems in this language pair were chosen for experimentation: the production of "tones" and the production cf aspirate and unaspirate voiceless initial stops.

Any isolated syllable in Chinese (a sequence of optional consonant and final vowel or vowels) can be pronounced with one of four tones, or movements of the fundamental frequency over the voiced portion. Depending on the tone used, the syllable's meaning changes. In the English transliteration, used in most American curricula, diacritical markings above the vowels indicate the tone to be used.

In multisyllabic utterances the contours of the tones associated with the component syllables may be modified. This is called "Tone Sandhi." An example is the "half-third-tone," a low and steady variant of the normally low-scooping isolated third tone. The half-third occurs in word-initial position. Another example is the "neutral" tone for unstressed syllables. It corresponds roughly to the "schwa" vowel in English. Its pitch contour is strongly dependent on the tone of the preceding stressed syllable. Relations between adjacent tones are often complex, and much drilling is required before the proper combinatory behavior is achieved.

The aspirate and unaspirate voiceless initial stops in Chinese differ from their counterparts in English. For example, the aspirate /p/ in "pill" is produced by emitting a puff or air (aspiration) prior to the onset of voicing. The corresponding Chinese aspirate initial, while it may be transliterated similarly, differs in that the puff of air is emitted with noticeably more force. The unaspirate opposite of /p/ is /b/, and in English this is produced by beginning voicing prior to or coincident with the opening of the lips, with the amount of air puffed at the moment of opening much smaller than /p/. The corresponding Chinese unaspirate initial begins the voicing in exact coincidence with the parting of the lips, with the intraoral pressure buildup at a minimum. The free variation in voice onset time for English but not Chinese may lead to confusion in the student between Chinese versions of /p/ and /b/.

There are four basic contrasts grouped as aspirate/unaspirate voiceless initials, depending on place of articulation. /p/ - /b/ was described above. The second, /t/ - /d/, is the labiodental contrast, with the emphasized aspiration of /t/ and the minimized aspirate, nonprevoiced /d/. The third is /g/ - /k/, glottal, with the /k/ produced in a manner easily confused by the student with the English /g/. The last contrast is transliterated "c - z," with no direct English equivalent. The "t's"-like sound of friction is emphasized in the "c" and it occurs before the voicing onset of the following vowel. In the "z" sound, voicing occurs earlier, but not before release.

2. METHOD

2.1 Selection and Pretesting

English speaking students of basic Mandarin Chinese were recruited from the introductory Mandarin Chinese courses at Harvard University and Massachusetts Institute of Technology. Brief presentations were made in regular classes to explain the purpose and pay scale of the experiment. All 14 volunteers were accepted into the study, half as experimentals, half as controls.

A test list of utterances was compiled to be administered to both groups three times. The first was a pretest given before training. The second was a post test immediately following the training of the experimentals, and the third a retention test after a no-treatment interval for both groups.

For each of these lists the students read a series of 24 two-syllable word pairs and phrases, under conditions controlled by a simple set of instructions read from the display screen of the API system. A tape recording was made of their speech. Table 1 gives the list of utterances produced. The four sections indicated on this list reflect the four segments of training administered to the experimental students. There were six utterances comprised of minimal pairs of single, isolated tones. This section thus tested production of unencumbered tone gestures. There were six disyllabic, two-tone utterances, testing for the proper combination of tones and including several words where tone sandhi radically alters the rendition of a component. The next six tested utterances were also disyllabic, but the second member was the so-called "neutral tone," The final six utterances were minimal pairs differing not in tone but in the initial stop.

Both groups were given the pretest. A teacher of Mandarin Chinese, who had recorded the API training tapes, listened to the tapes of their speech and rated all the utterances of all the volunteer subjects. An informal attempt was then made, with his help, to divide the subjects by proficiency equally into the experimental and control groups. Within each group there was a great deal of variance in pronunciation abilities.

2.2 Training

A curriculum was prepared in consultation with faculty teaching the Introductory Mandarin Chinese courses at Harvard

TABLE 1. TEST LIST

DISCRIMINATION		#	UTTERANCE
DISCRETE TONES	1-2	1.	MĪ MÍ
	3-1	2.	FǍ FĀ
	2-3	3.	MÁ MǍ
	1-4	4.	YĪ YÌ
	4-3	5.	YÀ YǍ
	4-2	6.	MÀO MÁO
2-TONE COMBINATIONS	1-2	7.	FĀ MÍNG
	4-1	8.	HÒU FĒI
	3-2	9.	MǍ FÁNG
	4-4	10.	YÀ LÀN
	2-3	11.	LÓNG YǍ
	3-3	12.	MǍ YǏ
VARIOUS TONES FOLLOWED BY NEUTRAL TONE	1-	13.	TĀ .LE
	3-	14.	YǍO .LE
	2-	15.	HÚ .LE
	3-	16.	TǍO .DE
	4-	17.	LÀ .DE
	4-	18.	YÀO .LE
ASPIRATED-UNASPIRATED VOICELESS INITIALS, VARIOUS TONES	B-P 1	19.	BĒI PĒI
	T-D 1	20.	TŪ DŪ
	T-D 4	21.	DÀ TÀ
	K-G 1	22.	KĀI GĀI
	Z-C 3	23.	ZÀO CǍO
	Z-C 1	24.	CĀI ZĀI

University and Massachusetts Institute of Technology. The goal of this effort was a set of materials that would supplement normal course work for the students. The same orthographic system as used in the students' textbooks was implemented on the API. The chosen subset of the pronunciation problems they faced was presented in the same manner as in the standard language laboratory materials available to all students. Since it was impossible to provide supplemental non-API training to the control group, it was important that they have access to similar materials in the parent course. The control group received no special treatment save the encouragement to utilize the language laboratory curriculum that was equally available to both experimenal and control students.

The seven experimental students each underwent eight training sessions on the API system. Each session involved from 35 to 45 minutes of training time without monitor intervention.

Sessions 1 and 2: Isolated identical tones. The first exposure of experimental students to the system was done with the simplest possible element of the curriculum. Each of the five tones was represented by four or five items in the 24-stimulus wordlist shown in Table 2. As in the parent course, the half-third tone was considered a separate entity in early training even though it never occurs in isolation. Each training utterance consists of two differing "carrier syllables" with the same tone on each syllable.

The speech function displayed was pitch. A few minutes of the first session were devoted to instruction in the manipulation of the equipment and in the interpretation of the display. The monitor soon left the students to their own devices and only needed to provide occasional further help.

The operation of the computer programs had been made flexible to allow variation in the possible approaches to different problems.

TABLE 2: TRAINING LIST 1: ISOLATED IDENTICAL TONES

DISCRIMINATION	#	UTTERANCE		FUNCTION DISPLAYED
TONE 1	1.	Ā	MĀ	PITCH: SLIDING MATCH, THEN VERTICAL PAIR MATCH
	2.	YĪ	FĒI	
	3.	TĀNG	FĀ	
	4.	YĀ	AĪ	
	5.	LĀO	MĀO	
TONE 2	6.	Á	MÁ	
	7.	YÁO	MÍ	
	8.	FÁN	MÁN	
	9.	YÍ	FÉI	
	10.	FÁ	YÁ	
1/2 TONE 3	11.	Ǎ-	MÁ-	
	12.	ǍI-	YÁ-	
	13.	FÈI-	YÌ-	
	14.	MǍO-	YǍO-	
TONE 4	15.	À	MÀ	
	16.	AÌ	LÀO	
	17.	FÀ	MÌ	
	18.	MÀN	WÀN	
	19.	FÈI	MÀO	
TONE 3	20.	À	MÀ	
	21.	YÌ	AÌ	
	22.	MÌ	FÈI	
	23.	LAO	MÀO	
	24.	YÀ	FÀ	

Many of the training sessions required different types of comparisons, and so the display procedures were altered to maximize the visual discriminability of the relevant parameters. The basic framework of the display, constant throughout, contained space for one or two teacher utterances and one or two student utterances. The student could match his utterances with those of the teacher, or could match his second with his first word, depending on what was being trained.

The major lesson to be learned in the first session was consistency in the production of tones. To aid the students, the software operated a "Match" function in "sliding mode." When the Match button was depressed, the second member of both the student's and teacher's pair of word traces described a smooth leftward motion until its first point met the first word's starting point. In the second training session "vertical pair" Match was used. While it was not strictly necessary in the context of the first training word list, it served as a simple introduction to the idea of inter-speaker comparison, used later. The two student word traces were each moved up smoothly until each one's starting point was at the same horizontal position as the corresponding teacher word. The student was instructed to attend to the parallelism between his trace and the teacher's, and to disregard absolute differences in fundamental frequency. The logarithmic nature of the pitch display facilitated this.

Sessions 3 and 4: Isolated tones in differing minimal pairs. Each of the possible tone pairs was presented, including pairs with the half-third tone as both the first and second member of the two-syllable utterance. Table 3 shows the word list. Since the two components of the minimal pair are pronounced as separate words, this utterance is not normal in spoken Mandarin, but again had been used in the parent course work. Two sessions were devoted to this wordlist. The first of them used sliding mode Match, so that the students could concentrate on producing the

TABLE 3: LIST 2: ISOLATED DIFFERENT TONES

DISCRIMINATION	#	UTTERANCE		FUNCTION DISPLAYED
1 - 2	1.	Ā	Á	PITCH: SLIDING MATCH, THEN VERTICAL PAIR MATCH
1 - 1/2 3	2.	MĀ	MǍ-	
1 - 3	3.	YĪ	YǏ	
1 - 4	4.	FĒI	FÈI	
2 - 1	5.	FÁ	FĀ	
2 - 1/2 3	6.	YÁ	YǍ-	
2 - 3	7.	ÁI	ǍI	
2 - 4	8.	LÁO	LÀO	
1/2 3 - 1	9.	FǍN-	FĀN	
1/2 3 - 2	10.	MǍN-	MÁN	
1/2 3 - 3	11.	MǏ-	MǏ	
1/2 3 - 4	12.	MǍO-	MÀO	
3 - 1	13.	Ǎ	Ā	
3 - 2	14.	MǍ	MÀ	
3 - 1/2 3	15.	YǏ	YǏ-	
3 - 4	16.	FĚI	FÈI	
4 - 1	17.	FÀ	FĀ	
4 - 2	18.	YÀ	YÁ	
4 - 1/2 3	19.	ÀI	ǍI-	
4 - 3	20.	LÀO	LǍO	
1 - 4	21.	FĀN	FÀN	
2 - 3	22.	MÁN	MǍN	
1 - 2	23.	MĪ	MÍ	
3 - 4	24.	MǍO	MÀO	

different tones in the proper pitch relations to each other. Session 4 addressed the problem of timing (tone duration) and used vertical-pair Match mode. Students could still make intra-speaker comparisons of trace shape as well, because they could compare their own two traces with the teacher's even before using the Match operation.

Sessions 5 and 6: Two-tone combinations. Tables 4 and 5 contain the two word lists, each of which taps many of the possible two-syllable tone combinations. Doubled tones are included since tone sandhi is often a factor. Difficult combinations, such as those involving special tones (such as half-third or half-falling fourth) are emphasized by repetition.

Students worked on the above two lists for one session each. The matching mode used for this material is called "vertical phrase," signifying that the entire student utterance is translated vertically without subdivision to superimpose on the teacher's entire utterance.

At this point in the training, the schedule underwent a forced modification. The experiment was being conducted during the Fall semester. The planned termination of the training sessions had been quite close to the Christmas holidays. However, earlier departures were unexpectedly planned by at least two experimental students, forcing the premature termination of the training for the entire group. A decision was made to train the last two wordlists with one session each, rather than abandon either entirely. Such are the limitations encountered in a semi-voluntary setting.

Session 7: Neutral tone following each of the four tones. This single training session used the wordlist shown in Table 6. A syllable written with no diacritical tone marker over its vowel and preceded by a period is pronounced with an unstressed neutral tone whose duration and contour depends on the preceding stressed tone. When the third tone precedes the neutral, its production shifts to the half-third. As in the other two-tone combinations, vertical phrase matching was used.

Session 8: Aspirate/unaspirate voiceless initial stops. The word list shown in Table 7 was used in the last training session. Each of the four contrasting consonant pairs is represented by a group of six minimal pair items in this list. Successive items reversed the direction of this discrimination; i.e., if one training item has the aspirate member of the pair first, the succeeding minimal pair will have the unaspirate member first. Tone within an item was constant, and an effort was made to have all tones represented in each of the four categories.

The display used for this material gave feedback principally on the presence and time course of speech noise produced <u>before</u> voicing onset. Both voice pitch and overall loudness of the speech were plotted as a composite during voiced sections of utterances: for voiced sections of speech, the familiar pitch trace appeared as before, but added above it was a set of dimmer points at a distance above the pitch trace proportional to the loudness of the voiced speech sound. Unvoiced speech sounds, which formerly (in earlier displays used by the students) had produced no visual feedback, now produced a single line near the bottom of the display at vertical positions proportional to the loudness of the unvoiced sound at that point in time. The distinction between voiced and unvoiced sounds was thereby made clear to the speaker, and he was to use the information in evaluating the relations between voiced

TABLE 4. LIST 3a: 2-TONE COMBINATIONS

DISCRIMINATION	#	UTTERANCE	FUNCTION DISPLAYED
1 - 1	1.	TĀ TĪNG	PITCH, VERTICAL PHRASE MATCH
1 - 2	2.	TĀ LÁI	
1 - 3	3.	TĀ MǍI	
1 - 4	4.	TĀ MÀI	
2 - 1	5.	MÉI TĪNG	
2 - 2	6.	MÉI LÁI	
2 - 3	7.	MÉI MǍI	
2 - 4	8.	MÉI MÀI	
3 - 1	9.	NǏ TĪNG	
3 - 1	10.	MǍ ĀN	
3 - 2	11.	NǏ LÁI	
3 - 2	12.	HǍO WÁNR	
3 - 3	13.	NÌ MǍI	
3 - 3	14.	LǍO HǓ	
3 - 3	15.	MĚI MǍN	
3 - 4	16.	NǏ MÀI	
3 - 4	17.	MǍ LÙ	
4 - 1	18.	YÀO TĪNG	
4 - 2	19.	YÀO LÁI	
4 - 3	20.	YÀO MǍI	
4 - 3	21.	TÀI LǍO	
4 - 4	22.	YÀO MÀI	
4 - 4	23.	HÀO WÀI	
4 - 4	24.	YÀO FÀN	

TABLE 5: LIST 3b: 2-TONE COMBINATIONS

DISCRIMINATION	#	UTTERANCE	FUNCTION DISPLAYED
1 - 1	1.	SĀN FĀN	PITCH: VERTICAL PHRASE MATCH
1 - 2	2.	YĀ LÚ	
1 - 3	3.	TĀ HǍN	
1 - 4	4.	FĀ LÌNG	
2 - 1	5.	LÍ MĀO	
2 - 2	6.	YÁO LÍNG	
2 - 3	7.	HÁI HǍO	
2 - 4	8.	FÚ LÌ	
3 - 1	9.	LǍO MĀO	
3 - 1	10.	TǏNG HĒI	
3 - 2	11.	MǍN TÁNG	
3 - 2	12.	LǏANG PÍNG	
3 - 3	13.	TǓ FĚI	
3 - 3	14.	MĚI MǍN	
3 - 3	15.	LǍO HǓ	
3 - 4	16.	LǏ FÀ	
3 - 4	17.	LÀO HÙA	
4 - 1	18.	TÀI HŪA	
4 - 2	19.	SÙ LÁI	
4 - 3	20.	FÙ MǍ	
4 - 3	21.	LÌ FǍ	
4 - 4	22.	MÙ TÀN	
4 - 4	23.	YÀO FÀN	
4 -4	24.	HÀO WÀI	

TABLE 6: LIST 4: NEUTRAL TONE

DISCRIMINATION	#	UTTERANCE	FUNCTION DISPLAYED
1	1.	TĪNG .LE	PITCH: VERTICAL PHRASE MATCH
	2.	SĀN .GE	
	3.	FĒI .DE	
2	4.	LÁI .LE	
	5.	YÍ .GE	
	6.	PÁ .DE	
1/2 3	7.	MǍI .LE	
	8.	WǓ .GE	
	9.	PǍO .DE	
4	10.	MÀI .LE	
	11.	LÌU .GE	
	12.	TIÀO .DE	
1	13.	MĀ .MA	
	14.	TĀ .DE	
	15.	FĀN .LE	
2	16.	LÍ .BA	
	17.	MÁI .LE	
	18.	LÁN .DE	
1/2 3	19.	HǍO .BA	
	20.	FÀN .DE	
	21.	SǍO .LE	
4	22.	HÀI .TA	
	23.	LÈI .LE	
	24.	SÙ .DE	

TABLE 7: LIST 5: ASPIRATE/UNASPIRATE VOICELESS INITIALS

DISCRIMINATION	tone	#	UTTERANCE	FUNCTION DISPLAYED
B-P	4	1.	BÈNG PÈNG	PITCH-LOUDNESS COMPOSITE: VERTICAL PAIR MATCH
	1	2.	PĀN BĀN	
	2	3.	BÁI PÁI	
	3	4.	PÀO BǍO	
	4	5.	BÙ PÙ	
	3	6.	PǏAO BǏAO	
D-T	4	7.	DÙI TÙI	
	1	8.	TĪ DĪ	
	1	9.	DĪNG TĪNG	
	3	10.	TǍO DǍO	
	4	11.	DÀNG TÀNG	
	3	12.	TÓNG DǑNG	
G-K	4	13.	GÀN KÀN	
	1	14.	KĀNG GĀNG	
	3	15.	GǓ KǓ	
	3	16.	KUǍI GUǍI	
	3	17.	GǑNG KǑNG	
	4	18.	KÀU GÀU	
Z-C	4	19.	ZÙI CÙI	
	1	20.	CĀI ZĀI	
	2	21.	ZÁO CÁO	
	3	22.	CǍN ZǍN	
	1	23.	ZŪ CŪ	
	1	24.	CĀNG ZĀNG	

and unvoiced consonants along the lines discussed in the preceding phonological introduction. Students reported little trouble in using the display for consonants, and some reported that its pitch feedback served as a good review for simple tone production they had studied previously.

2.3 Post- and Retention-Testing

Both groups of students were post-tested at roughly the same time. Both groups read the same list of 24 test utterances they had first seen at pretest time and following the same procedure. The material in the testing list had not appeared in the training wordlists for the experimental students, and it had been seen by both groups in the course of their normal language laboratory work.

2.4 Evaluation Procedures

Each student had recorded his best attempts at the 24 test utterances at three points in time. The test day tape recordings were copied, cut, and spliced such that a set of 14 judgment tapes, one for each of the students, was prepared. Each judgment tape began with the student reading two sample English sentences, to enable a listening judge to form some idea of the normal tone of the student's voice. Then followed four similar sections, based on each of the four segments of the test list. First, the six utterances as read by the native Mandarin teacher were heard. Then, separated from each other by approximately five seconds, the 18 versions of the six test utterances were heard in a scrambled order whose only constraint was that the same utterance's three versions could not be heard in three successive positions. No identification of student or of testing day was contained on the tape.

Five instructors of Introductory Mandarin Chinese from Boston area universities, all Mandarin natives, served as paid judges. Each judge worked alone in the API student room, listening to the 14 judgment tapes played over the student loudspeaker at a comfortable listening level. The order of students was unique for each judge, and was counterbalanced to compensate for increasing familiarity with the judgment task. Two 15-minute rest periods were interposed within the approximately 4-hour course of each judge's ratings.

Written instructions (included in Appendix 1) for the judges explained the rating scale they were to apply. Each test utterance was to be assigned an integer number from 0 to 4, higher numbers associated with better performance.

To aid them further in their task, each judge had a short-form rating instruction sheet (Appendix 2) and, for each judgment tape, an actual script of the order of the utterances (a sample is shown in Appendix 3). This "answer sheet" did not, of course, identify either student or test day, but it did serve to inform the judge of what test utterance the student was in fact attempting. This was particularly valuable in cases of gross student error. Three orthographic systems were used in identifying the test utterances on the judges' sheets, so that they could utilize the most familiar one. Judges wrote their accent ratings in a blank following each line of the answer sheet.

Judges could ask the assistant to stop or replay the tape to give them more time to come to a decision, but these requests diminished over time and the data were gathered without incident.

Since both groups of students were part of a larger-scope course in basic Mandarin Chinese, it was expected that their overall Chinese speech quality would improve through time, irrespective of their status within the experiment. The central question addressed to the data was, therefore, whether there was a differential improvement between the students using the API system and the group not.

The word lists used for testing were divided by the four types of training materials: separate single tones, disyllabic combinations, neutral tone disyllables, and consonant contrasts.

A measure of the student's performance for each of the four sections of the test lists, for each test day, over all judges was obtained. Then the data for all experimental subjects were combined and compared with all the controls.

A judgment is defined as the score given by one judge to one word spoken by one subject on one test day. Comparing, for example, pre and post judgments of one word, a subject could receive a higher post score, a lower post score or the same score. If he received a higher post score, he improved his pronunciation of that word from the pretest to the post test according to the judge. Two comparisons were made: pre vs. post tests and pre vs. retention tests.

3. RESULTS

Considering first the pre vs the post tests, over half the judgments made by all judges, for all the subjects and all the words showed no change in pronunciation ability. For the experimentals, 58 percent, and for the controls 62 percent of all judgments made on the pre-test words did not change on the post-test. Of the judgments that did change, the experimentals were more likely to have improved than the controls, while the controls were about equally likely to have scored lower as higher when changes occurred from pre to post tests. Controls improved 54 percent of the time when they changed, while the experimentals improved on 73 percent of all changed judgments. This difference is significant ($p<.001$). Table 8 gives more detail.

The pre- vs retention-test comparisons showed similar trends. The experimental subjects retained the improvements they had made on the post tests. The controls, who showed very

TABLE 8. PRE-POST TEST COMPARISONS
OVER ALL WORDS

	Experimentals		Controls	
	#	%	#	%
Total number of judgments indicating no change	487	58	524	62
Total number of judgments indicating improvement	257	31	170	20
Total number of judgments indicating poorer pronunciation	96	11	146	18

X^2 including only judgments indicating change = 26.09
df=1
$p<.001$

little average change from the pre to the post test improved their performance on retention. Of the judgments that did show a change, the experimentals improved on 73 percent and the controls 66 percent. There was still a large number of judgments, in both groups, that showed no change in performance from pre to retention tests, 56 percent of all judgments for the experimentals and 67 percent for the controls. See Table 9.

The distribution of "no change" judgments was even over all four word groups for experimentals and controls. See Table 10. The controls were not more or less likely than the experimentals to "not change" from pre to post tests or pre to retention tests. None of the four stimulus word groups was more or less likely than any other to show changed judgments.

The greatest differential improvement of experimentals over controls occurred on the first group of the stimulus list, the isolated single tones. This was the simplest element of the curriculum. The subjects had received four relevant sessions of training, for this type of material. Whether the differences in performance arise from the type or amount of training given or the nature of the stimulus material cannot be ascertained. However, significantly more of the judgments of improvement occurred among the experimentals rather than the controls. See Table 11. The differences between experimentals and controls on this group of tones also remained significant on the retention test. See Table 12.

The second and third test word groups were the disyllabic combinations and the neutral tone disyllables. The experimentals consistently had a greater number of judgments showing improvement than the controls, over both word groups and on the pre-post

TABLE 9. PRE-RETENTION TEST COMPARISONS
OVER ALL WORDS

	Experimentals		Controls	
	#	%	#	%
Total number of judgments indicating no change	469	56	563	67
Total number of judgments indicating improvement	272	32	184	22
Total number of judgments indicating poorer pronunciation	99	12	93	11

X^2 including only judgments indicating change = 3.61
df=1
$p<.10$

TABLE 10. DISTRIBUTION OF "NO CHANGE" JUDGMENTS

PRE-POST COMPARISONS

	WORD GROUP 1		WORD GROUP 2		WORD GROUP 3		WORD GROUP 4	
	# no change judgments	% of all judgments	# no change judgments	% of all judgments	# no change judgments	% of all judgments	# no change judgments	% of all judgments
Experimentals	117	24	108	22	128	26	134	28
Controls	140	27	125	24	130	25	129	24
All Subjects	257	25	233	23	258	26	263	26

PRE-RETENTION COMPARISONS

	WORD GROUP 1		WORD GROUP 2		WORD GROUP 3		WORD GROUP 4	
	# no change judgments	% of all judgments	# no change judgments	% of all judgments	# no change judgments	% of all judgments	# no change judgments	% of all judgments
Experimentals	119	25	102	22	120	26	128	27
Controls	152	27	127	22	140	25	144	26
All Subjects	271	26	229	22	260	25	272	27

TABLE 11. PRE-POST TEST COMPARISONS BY WORD GROUP

WORD GROUP 1

	Experimentals	Controls
Total number of judgments indicating improvement	61	23
Total number of judgments indicating poorer pronunciation	32	47

$X^2=17.13$ df=1 $p<.001$

WORD GROUP 2

	Experimentals	Controls
Total number of judgments indicating improvement	78	57
Total number of judgments indicating poorer pronunciation	24	28

$X^2=2.05$ df=1 $p<.25$

WORD GROUP 3

	Experimentals	Controls
Total number of judgments indicating improvement	56	46
Total number of judgments indicating poorer pronunciation	26	34

$X^2=2.02$ df=1 $p<.25$

WORD GROUP 4

	Experimentals	Controls
Total number of judgments indicating improvement	62	44
Total number of judgments indicating poorer pronunciation	14	37

$X^2=13.28$ df=1 $p<.001$

TABLE 12. PRE-RETENTION TEST COMPARISONS BY WORD GROUP

WORD GROUP 1

	Experimentals	Controls
Total number of judgments indicating improvement	70	24
Total number of judgments indicating poorer pronunciation	21	47

$X^2=30.45$ df=1 $p<.001$

WORD GROUP 2

	Experimentals	Controls
Total number of judgments indicating improvement	85	60
Total number of judgments indicating poorer pronunciation	25	23

$X^2=.63$ df=1

WORD GROUP 3

	Experimentals	Controls
Total number of judgments indicating improvement	58	42
Total number of judgments indicating poorer pronunciation	31	31

$X^2=.99$ df=1

WORD GROUP 4

	Experimentals	Controls
Total number of judgments indicating improvement	59	47
Total number of judgments indicating poorer pronunciation	22	16

$X^2=.06$ df=1

and pre-retention comparisons. The contrast between the experimentals and the controls was not as great as on the first word group, however. Only three training sessions were given to these tone combinations, both using inter-speaker comparisons. The rate of improvement of the experimentals was about the same as on the first word group; the controls showed more important than they had on the single tones, and the differences between the groups were not as great.

The fourth test word group consisted of consonant contrasts. On this set of words the experimentals improved significantly over the controls on the pre to post test but not on the pre-retention test comparison. Of the judgments that did indicate change, the experimentals improved on 82 percent of the post-test judgments compared with 54 percent for the controls, and on 73 percent of the retention judgments compared with 75 percent of the controls. Aspirate and unaspirate voiceless initial stops could only be trained for one session, with the more complex pitch-loudness composite display.

4. DISCUSSION

Despite the severe limits in the breadth of the student sample and in the time available for training, a real improvement was generally observed in the Chinese speech of the students exposed to the API system, an improvement significantly greater than that observed in students tested similarly but exposed only to the "parent" Chinese course. One must keep the limitations of the present experiment in mind when assessing the performance of the API system in this situation. Though the effects observed were small statistically, they are nonetheless real, and their size is probably limited more by the scope of the work than by the efficacy of the system. To have observed significant treatment effects in the face of short training time and an inherently "noisy" evaluation procedure speaks strongly for the robustness of that treatment effect.

The meaning of the treatment effect should be evaluated in light of two opposed factors. On the one hand, the test list was drawn from parent course materials, so that both experimental and control students would have the same basic familiarity with the utterances. Furthermore, materials tested had not been included within the training materials used by the experimental students. Any observed treatment effects can thus be ascribed to differential pronunciation ability rather than to increased familiarity with the testing utterances. On the other hand, the sample of speech behavior obtained from the students intentionally included only utterances of a type similar to those trained, so that any possible treatment effects would stand out in sharp relief.

One consequence of the limited scope of the speech behavior tested is the restriction on the inferences that may be drawn concerning the overall pronunciation abilities of the experimental subjects. This was done with the realization that the most sensitive means of evaluation could be applied only to speech behaviors easily judged and reliably produced. The primary hurdle the API must pass is a demonstration that it can produce improvements in accent, but it is unrealistic to expect either that (a) training on a specific set of accent problems will produce an across-the-board improvement in pronunciation, or (b) that a panel of accent-rating judges can make reliable responses concerning anything as multidimensional as "total accentedness" of a set of utterances. The evaluation method chosen, and the statistical procedure used to reduce the data, were therefore designed to produce maximal sensitivity to change while at the same time avoiding the more complex method of complete pair-comparisons. A single-stimulus rating technique by a panel of judges produced responses that could be subjected to a pair-comparison-type analysis, if due regard were given to the permissible operations on the data. As it happens, one is in fact interested not in comparisons between specific words and subjects, but in accent parameters (i.e., specific word groups), treatments (i.e., experimental or control), and testing times (pre-, post-, or retention-testing data). The present analysis provides answers to questions posed along those lines, having minimized the variance produced by both the speech production and subjective judgment processes.

The major price paid in the analysis is the large number of "no change" judgments encountered. These result largely from the coarse grain of the judgment scale. Taking this price into account, one is still left with a reasonable statement of the null hypothesis as regards the treatment effect: that there is no difference between treatment groups in the distribution of "improved" versus "poorer" pronunciations. That hypothesis fails of acceptance in a consistent manner throughout the above analysis.

Tables 8 and 9 showed that the number of equivocal judgments for all test words was smaller for experimentals than for controls, in both pre-post and pre-retention comparisons. Furthermore, it was shown that when there was a change, it was significantly more often in the direction of improvement for the experimentals than for the controls; they learned more and retained it better.

After having been assured by Table 10 that the equivocal judgments distribute themselves evenly across the four word groups, it becomes reasonable to inspect individual word groups' changed judgments for differences in distribution as a function of treatment. Again, it is fcund (in Tables 11 and 12) that in each word group and for both pre-post and pre-retention comparisons, the experimentals' changes are always in the direction of greater improvement, and significantly so in three out of the eight specific comparisons made. The strong showing made by word group 1 is not surprising; it received the largest share of the training time, and was conceptually the simplest display. The unexpectedly strong treatment effect observed in word group 4 is most easily explained by the action of the pitch-loudness composite display used there. Even though the training time available to the experimentals for this work group was but one session, they apparently profited greatly from even this brief exposure to the display. Since all observed effects favored the experimental treatment, it is reasonable to take the position that a simple increase in training time might have brought all differential treatment effects to significant levels.

At this writing, the final field tests of the API system are underway at the University of Miami's Intensive English Program, Coral Gables, Florida, with Spanish speakers learning English. This experimentation is much broader in scale. Experimental variables are under better control in that situation, and the scope of problems trained and measurements taken is larger. The work reported above gives reason for optimism, because even when the system is tested under less than optimal conditions, significant benefits accrue to its students. Subsequent reports in this series will describe the results of a field test in which the API is used as a part of the daily schedule of a group of second-language learning students.

APPENDIX 1

INSTRUCTIONS TO JUDGES

INSTRUCTIONS TO JUDGES

Your task today is to evaluate Chinese utterances made by students of Introductory Mandarin Chinese, who were also subjects in an experiment designed to test a Chinese pronunciation teaching-machine. Each student read a set of test words at various times throughout the experiment. We wish to find out whether the students' pronunciation of those test words changed over time. The utterances have been randomly scrambled and collected onto "judgment tapes," one judgment tape for each student. You will sit alone in a listening room and you will assign a numerical grade to each utterance as you hear it. The tape contains adequate time for you to consider and respond to each item, before the next one is heard. If you need additional time, or if you want to pause for any reason, there is a microphone connected outside, enabling you to ask the operator to wait. When you are ready to resume, tell him and things will proceed.

There are two booklets to aid you in assigning the grades to the students' utterances. The small, four-page booklet is the key to what the utterances are, and to how the grading is to be made. .Each page corresponds to one of the four sections of the tape from each student. Each section deals with six words or word pairs. The <u>bottom half</u> of each page contains transcribed English and two Chinese script versions of the six words that have been scrambled up three times to form one 18-utterance section of each student's tape. The <u>top half</u> of each page gives a brief synopsis of the grading scheme for each section. (The last part of these instructions will give you detailed information on how to grade each section's utterances; for now, let us assume that you will, in general, assign each utterance a grade ranging from 0 to 4, bad to good, in accordance with the instructions and with your judgment.)

The thicker booklet is your key to the utterances themselves. It is, in essence, a script that tells you what word(s) the student was actually attempting to produce. It will help you keep your place. It gives you a blank space within which you are to write your judgment of each utterance. It will be especially helpful when the student's version of the intended utterance is garbled. By knowing what the student was trying to say, you can better judge how well he succeeded. Make sure that each line receives a written response from you -- either 0, 1, 2, 3, or 4. If you need more time to consider your judgment, just ask for a pause. If you would like to hear any utterance over again, just ask for it.

Here is a view of what the judgment procedure is for the entire session. There will be 15 judgment tapes played. There will be a short break between tapes. Each tape has the same format as the others. The first voice you hear will not be that of the student whose utterances are collected on the tape; it will be an identifier for the tape number. Make sure that it corresponds to the tape number written on the top of the next sheet of the judgment booklet. If it does not, tell the operator, because the script will then not agree with the words you hear. At the start, then, the first page of the judgment booklet corresponds to the first section of the first tape.

After you have correctly identified the tape number and assured that your judgment booklet is on the right page, you will hear the student for the first time. He will read two sentences: "Joe took father's shoe bench out." and "She was waiting at my lawn." These sentences are merely for the purpose of acquainting you with the voice of the student before each tape actually begins. Through these introductory sentences, you can form an impression of his or her normal tone of voice, so that abnormal tone range will be apparent from the first time it appears.

Each judgment tape then continues with the four sections of 18 scrambled utterances of the student. For the first few judgment tapes, the operator will precede each 18-utterance section with a recording of a Mandarin speaker pronouncing the six utterances in the order given on the bottom of the four-page booklet. This is to familiarize you with the timing of the utterances, and to give you an example of the type of pronunciation that the students were attempting to imitate. As you become more experienced in listening to these tapes, you will have less need to hear the introductory Mandarin-native introduction to each of the four sections, and the operator will skip over it. If you want to hear it, just ask. At the end of the last teacher-version, there is a 10-second pause, and then the 18 utterances of the student will be heard. You will respond to each of them by placing a number in the appropriate blank of the answer sheet for that tape and section.

And now: What do those numbers mean? How are you to decide? First, remember that you are a native speaker of Mandarin, and you will have an instant opinion of each of the utterances, as to how they compare to your internal standard. Your teaching experience, and some knowledge of the mechanism of speech production -- especially for tones -- will also help you a great deal in assigning judgments. The utterances you will judge are quite short, which makes your job easier since there are fewer aspects of each utterance that you need to consider in making your judgment. Also, we are asking you to disregard certain irrelevant aspects of the students' speech, since they were only trained in the production of (in sections 1, 2, and 3) proper tones and (in section 4) proper initial aspirate and unaspirate stops. The top line of the four-page handout indicates what was trained (i.e., what to pay attention to) and what to disregard in making your judgments.

SECTION I: Separate tones

In this section, as in all the rest, if the student's utterance (for the appropriate aspects) is OK, score it 4. If it is less than OK, think of the following breakdown of his performance. There are two words, each with two aspects: duration (total time for the tone) and contour (voice pitch as a function of time). While they are not really separable, try to make them so for the present purpose. The two tones are also produced with a given relative pitch level. If you can pinpoint just one error in the two word utterance, score it 3. Possible errors, then: (a) one tone too short or too long, (b) one contour off slightly, (c) both tones OK but relative pitch wrong, (d) "just slightly off -- and definitely not OK," etc. These might be 3's. A score of 2 would be as indicated in the handout, and the remaining grades are self-explanatory. If the preceding sounds too complicated, remember the general idea and assign the grades from 0 to 4 according to the left-hand side of the grading description on the handout: 4 for OK, unaccented and 0 for unacceptable, a total miss.

Remember that the two words, while spoken together, are not really part of a complete two-syllable utterance. There only point of relationship is in their relative levels. The amount of time the speaker pauses between words is irrelevant.

SECTION II: Two-syllable tone pairs

Here, the two syllables are supposed to be pronounced together, and the linkage between them is a subject for scrutiny. The durations and contours of the two are important, the manner of their linkage is important, and the existence of tone sandhi is very important. Again, disregard all aspects of the utterances except the tones.

A rating of 4 signifies that the utterance is OK, unaccented. Give a 3 when it is "almost OK," but do not count as 3's any attempt that lacks the proper sandhi (influence by syllable 2 on syllable 1's tone structure). Give a 2 to utterances where there are two errors, and reserve 1 for sounds that are "better than nothing" or which are two tones lacking proper sandhi when appropriate. Give 0 to bad tries. As before, the general ordering from "4- OK" through "0 - bad" is an alternative mode of consideration for the judgments in this section.

SECTION III: Neutral tone as second member of two-syllable tone pairs

Use the same general approach as in Section II. The second syllable, the neutral tone, is short and doesn't have much contour, but its linkage to syllable one, and its sandhi upon syllable one, are of great interest.

SECTION IV: Aspirated and unaspirated initial consonants

Here, you are to try to disregard vowels and tones, and concentrate your attention on how well the speaker produces the consonants. The six word pairs alternate in which member of the pair is aspirated and which not. Each initial consonant has two general aspects: Voice-onset time and voice quality. The aspirate stops should exhibit the right sound of friction for the right amount of time before the vowel begins. The unaspirate stops should have a far shorter period of friction before the vowel, and they too should sound correct during the consonant portion. As you know, unaspirate stops must not be prevoiced in Mandarin. Follow the handout in assigning grades to these utterances. For example, give a 3 to a word pair where one word is OK and the other has one of the above errors.

GENERAL COMMENTS:

We realize that we are asking you to do a difficult task. We realize further that your grades may change over time. The purpose of the above standards is to provide you with some sort of absolute yardstick, but invariability is hard to come by in human judgments. We realize this too, and have allowed for it; so just try to do as well and as consistently as you can.

We expect that you will work as carefully and as conscientiously as possible. Much hangs in the balance in this experiment, and so we wish you to consider your judgments as carefully as possible within the time available. Remember that you are being paid about 5¢ per judgment, and try to provide your full attention to each utterance, disregarding any extraneous sounds that may have remained on the judgment tapes.

There will be speakers whose performance is better than others. Try not to let your scale become relative only to the present speaker, sliding up and down to match the level of each speaker. Try to remain unmoved by swings in ability, but to judge each speaker and indeed each utterance as an independent event. Your increasing experience in this judgment situation may cause some shifts through the entire session; don't become overly concerned with this. If you follow the general guidelines, that is enough for our purposes. Don't try to artificially distinguish between performances that are only slightly different. The categories are fairly broad, and a given level of grading can encompass utterances that differ.

What we are saying is: Try your best to give us a frank impression of how well each speaker produces each utterance -- the better the performance, the higher the score. If you follow the strategy outlined above, we will be satisfied.

BY ALL MEANS ASK ANY QUESTIONS YOU WISH, NOW OR AT ANY TIME DURING THE SESSION.

APPENDIX 2

SHORT-FORM RATING INSTRUCTION SHEET

Section I: Separate tones (disregard vowels and consonants)

WORD ONE

Duration

Contour

Relative Level

WORD TWO

Duration

Contour

4:	Unaccented	4:	All above points
3:		3:	One error above
2:		2:	"Half-credit"; One error in one tone, relative level wrong.
1:	↓	1:	"Better than nothing;" One word OK, the other wrong.
0:	Unacceptable	0:	Total miss

1.	mī	mí	咪 ㄇㄧ	謎 ㄇㄧˊ
2.	fǎ	fā	法 ㄈㄚˇ	發 ㄈㄚ
3.	má	mǎ	麻 ㄇㄚˊ	馬 ㄇㄚˇ
4.	yī	yì	衣 ㄧ	億 ㄧˋ
5.	yà	yǎ	壓 ㄧㄚˋ	啞 ㄧㄚˇ
6.	mào	máo	貌 ㄇㄠ	毛 ㄇㄠˊ

Section II: Two-syllable tone pairs (disregard vowels and consonants)

Syllable One		Syllable Two	
Duration		Duration	
	Relative Level		
Contour		Contour	
(Proper influence of Syllable Two)			

4:	Unaccented	4:	All above points
3:		3:	"Almost OK;" One error above, except proper "two-on-one" influence.
2:		2:	"Half-credit"
1:	↓	1:	"Better than nothing;" e.g.: no "two-on-one" influence, etc.
0:	Unacceptable	0:	Nothing

7. fā-míng 發ㄈㄚ 明ㄇㄧㄥˊ

8. hòu-fēi 後ㄏㄡˋ 妃ㄈㄟ

9. mǎ-fáng 馬ㄇㄚˇ 房ㄈㄤˊ

10. yà-làn 壓ㄧㄚˋ 爛ㄌㄢˋ

11. lóng-yǎ 聾ㄌㄨㄥˊ 啞ㄧㄚˇ

12. mǎ-yǐ 螞ㄇㄚˇ 蟻ㄧˇ

Section III: Neutral tone as second member of two-syllable tone pairs (disregard vowels and consonants)

Syllable One		Syllable Two
Duration	Relative Level	Duration
Contour		Contour
(Proper influence of Syllable Two)		

4:	Unaccented	4:	All above points
3:	↓	3:	"Almost OK;" One error above, exceot proper "two-on-one" influence.
2:		2:	"Half-credit"
1:		1:	"Better than nothing"
0:	Unacceptable	0:	Nothing

13.	tā .le	塌ㄊㄚ了ㄌㄜ˙
14.	yǎo .le	咬ㄧㄠˇ了ㄌㄜ˙
15.	hú .le	糊ㄏㄨˊ了ㄌㄜ˙
16.	tǎo .de	討ㄊㄠˇ的ㄉㄜ˙
17.	là .de	辣ㄌㄚˋ的ㄉㄜ˙
18.	yào .le	要ㄧㄠˋ了ㄌㄜ˙

Section IV: Aspirated and unaspirated initial consonants (Disregard vowels and tones)

ASPIRATED WORD (P,T,K)		UNASPIRATED WORD (B,D,G)	
Proper voice-onset time		No prevoicing, proper voice-onset time	
Proper spectral quality		Proper spectral quality	
4:	Unaccented	4:	All above points
3:	↓	3:	One word not quite OK
2:		2:	"Half-credit;" both not quite OK, or one word wrong
1:		1:	"Better than nothing"
0:	Unacceptable	0:	Nothing

19.	bēi	pēi	悲ㄅㄟ	胚ㄆㄟ
20.	tū	dū	突ㄊㄨ	都ㄉㄨ
21.	dà	tà	大ㄉㄚˋ	遢ㄊㄚˋ
22.	kāi	gāi	開ㄎㄞ	該ㄍㄞ
23.	zǎo	cǎo	早ㄗㄠˇ	草ㄘㄠˇ
24.	cāi	zāi	猜ㄘㄞ	災ㄗㄞ

APPENDIX 3

SAMPLE SCRIPT GIVING THE ORDER OF UTTERANCES IN ONE SUBJECT'S JUDGMENT TAPE

BBN SECOND-LANGUAGE PROJECT
CHINESE EVALUATION TEST DATA SHEET

TAPE NO.: 20 SECTION: 1 JUDGE: DATE:

4	yī	yì	衣	ㄧ	億	ㄧˋ	
2	fǎ	fā	法	ㄈㄚˇ	發	ㄈㄚ	
1	mī	mí	咪	ㄇㄧ	謎	ㄇㄧˊ	
5	yà	yǎ	壓	ㄧㄚˋ	啞	ㄧㄚˇ	
2	fǎ	fā	法	ㄈㄚˇ	發	ㄈㄚ	
4	yī	yì	衣	ㄧ	億	ㄧˋ	
6	mào	máo	貌	ㄇㄠˋ	毛	ㄇㄠˊ	
2	fǎ	fā	法	ㄈㄚˇ	發	ㄈㄚ	
4	yī	yì	衣	ㄧ	億	ㄧˋ	
1	mī	mí	咪	ㄇㄧ	謎	ㄇㄧˊ	
5	yà	yǎ	壓	ㄧㄚˋ	啞	ㄧㄚˇ	
3	má	mǎ	麻	ㄇㄚˊ	馬	ㄇㄚˇ	
5	yà	yǎ	壓	ㄧㄚˋ	啞	ㄧㄚˇ	
6	mào	máo	貌	ㄇㄠˋ	毛	ㄇㄠˊ	
6	mào	máo	貌	ㄇㄠˋ	毛	ㄇㄠˊ	
3	má	mǎ	麻	ㄇㄚˊ	馬	ㄇㄚˇ	
1	mī	mí	咪	ㄇㄧ	謎	ㄇㄧˊ	
3	má	mǎ	麻	ㄇㄚˊ	馬	ㄇㄚˇ	

BBN SECOND-LANGUAGE PROJECT
CHINESE EVALUATION TEST DATA SHEET

TAPE NO.: 20 SECTION: 2 JUDGE: DATE:

8	hòu-fēi	後ㄏㄡˋ妃ㄈㄟ	____
7	fā-míng	發ㄈㄚ 明ㄇㄧㄥˊ	____
8	hòu-fēi	後ㄏㄡˋ妃ㄈㄟ	____
11	lóng-yǎ	聾ㄌㄨㄥˊ啞ㄧㄚˇ	____
10	yà-làn	壓ㄧㄚˋ爛ㄌㄢˋ	____
11	lóng-yǎ	聾ㄌㄨㄥˊ啞ㄧㄚˇ	____
10	yà-làn	壓ㄧㄚˋ爛ㄌㄢˋ	____
12	mǎ-yǐ	螞ㄇㄚˇ蟻ㄧˇ	____
7	fā-míng	發ㄈㄚ 明ㄇㄧㄥˊ	____
12	mǎ-yǐ	螞ㄇㄚˇ蟻ㄧˇ	____
9	mǎ-fáng	馬ㄇㄚˇ房ㄈㄤˊ	____
9	mǎ-fáng	馬ㄇㄚˇ房ㄈㄤˊ	____
10	yà-làn	壓ㄧㄚˋ爛ㄌㄢˋ	____
7	fā-míng	發ㄈㄚ 明ㄇㄧㄥˊ	____
9	mǎ-fáng	馬ㄇㄚˇ房ㄈㄤˊ	____
8	hòu-fēi	後ㄏㄡˋ妃ㄈㄟ	____
11	lóng-yǎ	聾ㄌㄨㄥˊ啞ㄧㄚˇ	____
12	mǎ-yǐ	螞ㄇㄚˇ蟻ㄧˇ	____

BBN SECOND-LANGUAGE PROJECT
CHINESE EVALUATION TEST DATA SHEET

TAPE NO.: 20 SECTION: 3 JUDGE: DATE:

17	là	.de	辣ㄌㄚˋ的˙ㄉㄜ	____
18	yào	.le	要ㄧㄠˋ了˙ㄌㄜ	____
14	yǎo	.le	咬ㄧㄠˇ了˙ㄌㄜ	____
16	tǎo	.de	討ㄊㄠˇ的˙ㄉㄜ	____
16	tǎo	.de	討ㄊㄠˇ的˙ㄉㄜ	____
14	yǎo	.le	咬ㄧㄠˇ了˙ㄌㄜ	____
15	hú	.le	糊ㄏㄨˊ了˙ㄌㄜ	____
13	tā	.le	塌ㄊㄚ了˙ㄌㄜ	____
14	yǎo	.le	咬ㄧㄠˇ了˙ㄌㄜ	____
16	tǎo	.de	討ㄊㄠˇ的˙ㄉㄜ	____
15	hú	.le	糊ㄏㄨˊ了˙ㄌㄜ	____
18	yào	.le	要ㄧㄠˋ了˙ㄌㄜ	____
13	tā	.le	塌ㄊㄚ了˙ㄌㄜ	____
18	yào	.le	要ㄧㄠˋ了˙ㄌㄜ	____
13	tā	.le	塌ㄊㄚ了˙ㄌㄜ	____
15	hú	.le	糊ㄏㄨˊ了˙ㄌㄜ	____
17	là	.de	辣ㄌㄚˋ的˙ㄉㄜ	____
17	là	.de	辣ㄌㄚˋ的˙ㄉㄜ	____

BBN SECOND-LANGUAGE PROJECT
CHINESE EVALUATION TEST DATA SHEET

TAPE NO.: 20 SECTION: 4 JUDGE: DATE:

19	bēi	pēi	悲ㄅㄟ	胚ㄆㄟ	____
21	dà	tà	大ㄉㄚˋ	踏ㄊㄚˋ	____
24	cāi	zāi	猜ㄘㄞ	災ㄗㄞ	____
19	bēi	pēi	悲ㄅㄟ	胚ㄆㄟ	____
20	tū	dū	突ㄊㄨ	都ㄉㄨ	____
23	zǎo	cǎo	早ㄗㄠˇ	草ㄘㄠˇ	____
21	dà	tà	大ㄉㄚˋ	踏ㄊㄚˋ	____
24	cāi	zāi	猜ㄘㄞ	災ㄗㄞ	____
19	bēi	pēi	悲ㄅㄟ	胚ㄆㄟ	____
20	tū	dū	突ㄊㄨ	都ㄉㄨ	____
21	dà	tà	大ㄉㄚˋ	踏ㄊㄚˋ	____
23	zǎo	cǎo	早ㄗㄠˇ	草ㄘㄠˇ	____
22	kāi	gāi	開ㄎㄞ	該ㄍㄞ	____
23	zǎo	cǎo	早ㄗㄠˇ	草ㄘㄠˇ	____
22	kāi	gāi	開ㄎㄞ	該ㄍㄞ	____
24	cāi	zāi	猜ㄘㄞ	災ㄗㄞ	____
22	kāi	gāi	開ㄎㄞ	該ㄍㄞ	____
20	tū	dū	突ㄊㄨ	都ㄉㄨ	____